Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-13 21:46:30
REFERENCE NO: 181

Author Names & Affiliations

- Robert Hamers - University of Wisconsin-Madison

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Chemistry; Physical Chemistry, Analytical Chemistry, Chemical Imaging

Title of Submission

Infrastructure for integration of heterogeneous data sets.

Abstract (maximum ~200 words).

Chemistry has an inherent complexity in which scientific insights often arise from the ability to synthesize information from many different types of highly heterogeneous data sets. Even when data from many different instruments can be obtained, there are substantial barriers to effective integration and mining of these data associated with their heterogeneous nature and the fact that in many cases data formats are controlled by instrument manufacturers and open-source data standards do not exist or are not implemented. Significant advances in the effective utilization of existing data could be obtained through the development of open-source data standards along with appropriate metadata. Incorporation of such data into a nation-wide database could markedly accelerate the rate of scientific progress and reduce unnecessary replication of data that has been independently validated.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Chemistry research involves extremely diverse types of datasets. These datasets are highly heterogeneous in size and information content. In most cases, data sets reside within individual research laboratories and are not shared with outside users unless specifically requested. In some fields, such as crystallography, open-source data standards have emerged that provide opportunities for extensive data mining. In chemistry, experimental measurements such as NMR, optical spectroscopy, chemical imaging, x-ray and electron spectroscopies exist but in most cases instrument manufacturers have developed proprietary file formats that strongly limit the ability to glean new insights by establishing correlations between distinct data sets. The lack of open-source data standards with associated meta-data leads to large amounts of data being sequestered in individual laboratories with limited accessibility to other users and no way to use advanced cyber-infrastructure to establish links between different data sets. I believe the chemistry community would benefit tremendously from a large-scale effort to develop nationally accepted, open-source data standards and establish a national data repository that would allow

researchers to place data into the repository for open access by other researchers. This would provide the opportunity for data scientists to develop advanced software to seek out correlations between data sets and ultimately provide new scientific insights using already-existing data.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

This would require software infrastructure and development of data standards that would be accepted across the chemistry community.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

To be effective, there would need to be a reason for instrument manufacturers to incorporate open-source data standards into their software. This would likely require a strong motivation to do so. This could come, for example, if federal funding agencies required that instruments purchased with federal funds meet certain open-source standards developed by and accepted by the scientific community.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."